Kernel Flow Tutorial

Badr + Edmilson + Lu + Zheng

June 2022

Per Lu's suggestion, we are starting this document in the hopes that our reading group can result in a tutorial for Kernel Flow method, based on the 2019 Owhadi-Yoo paper [OY19] and [OS19], with the background theory filled in as necessary.

Contents

1 problem formulation

Problem 1:

Given input/output data $(x_1, y_1), ..., (x_N, y_N) \in \mathcal{X} \times \mathcal{Y}$. Recover an unknown function $u^\dagger$ mapping $\mathcal{X}$ to $\mathcal{Y}$ such as:

$$u^\dagger(x_i) = y_i, \forall i \in 1, ..., N$$

The goal of this project is learning a function from a finite number of sampled data points by

2 Reproducing Kernel Hilbert Space

ttp://songcy.net/posts/story-of-basis-and-kernel-part-2/

3 Optimal Recovery

3.1 Setting

$(\mathcal{B}, \|\cdot\|)$ be a space of functions that is a Banach space whose norm is given by

$$\|u\|^2 := [Q^{-1}u, u],$$

where $Q : \mathcal{B}^* \to \mathcal{B}$ is a bijective mapping from the dual space $\mathcal{B}^*$ and is

- symmetric: $[\phi, Q\varphi] = [\varphi, Q\phi]$;
- positive: $[\phi, Q\phi] \geqslant 0$,

for $\varphi, \phi \in \mathcal{B}^*$.

[Zheng] (i) need linear space $\mathcal{Y}$ so that the space $\mathcal{B}$ of functions $u : \mathcal{X} \to \mathcal{Y}$ can have linear structure. (ii) does Q need to be a linear isomorphism, i.e., continuous, linear with a continuous linear inverse? L^2 is isomorphic to its dual and so admits such a Q, false for general Banach space, but for general Hilbert space? (iii) given $\mathcal{B}$, is it easy to construct such a Q?

One can not directly compute $u^\dagger \in \mathcal{B}$ but only with a finite number of features of $u^\dagger$. For this reason, we introduce the information map $\Phi : \mathcal{B} \to \mathbb{R}^m$ given by

$$u \mapsto \Phi(u) = \big([\phi_1, u], \ldots, [\phi_m, u]\big),$$

where $\{\phi_i\}_{i=1}^m \subset \mathcal{B}^*$ are linearly independent set of functions. A solution operator is a possibly nonlinear map $\Psi : \mathbb{R}^m \to \mathcal{B}$ that uses only the values of the information operator. For any solution operator Ψ and any state $u \in \mathcal{B}$, the relative error corresponding to the recovery problem can be written

$$\mathcal{E}(\Psi, u) = \frac{\|u - \Psi(\Phi(u))\|^2}{\|u\|^2}.$$

An optimal recovery solution mapping $\Psi^* : \mathbb{R}^m \to \mathcal{B}$ is given by

$$\Psi^*(y) = \arg\min_{\Psi} \max_{u \in \mathcal{B}} \mathcal{E}(\Psi, u).$$

[Zheng] Intuitive interpretation: $\max_{u\in\mathcal{B}} \mathcal{E}(\Psi, u)$ gives the worst possible scenario for each given Ψ, then minimizing over all possible Ψ yields the best option Ψ^*. This is similar to the distance from compact set A to compact set B given by $\mathrm{dist}(A, B) := \min_{a\in A} \max_{b\in B} \mathrm{dist}(a, b)$.

We aim to find an explicit formula for the solution mapping $\Psi^*(y)$;

3.2 Projection properties

Note two definitions that will be used throughout the next few pages:

- The inner product in $\mathcal{B}$ is given by:

$$\langle u_1, u_2 \rangle = [Q^{-1}u_1, u_2]. \tag{1}$$

- The inner product in $\mathcal{B}^*$ is given by: $\langle \phi_1, \phi_2 \rangle_* = [\phi_1, Q\phi_2]$;

Moreover, observe that above definitions can be combined and we verify that:

$$\begin{aligned} \langle Q\phi_1, u_2 \rangle &= [Q^{-1}(Q\phi_1), u_2] \\ &= [\phi_1, u_2]. \end{aligned} \tag{2}$$

Let us denote the finite set of linearly independent functions by $\{\phi_1, \ldots, \phi_m\} \subset \mathcal{B}^*$ and its span by $\mathcal{L}$. We define the Gram matrix by

$$\Theta_{ij} = [\phi_i, Q\phi_j], \quad i, j = 1, \ldots, m,$$

and the elements $\psi_i \in \mathcal{B}$ by

$$\psi_i = \sum_{j=1}^{m} (\Theta^{-1})_{ij} Q\phi_j, \quad i = 1, \dots, m, \tag{3}$$

where $(\Theta^{-1})_{ij}$ denote the components of the inverse matrix Θ^{-1}.

[Lu] $(\Theta^{-1})_{ij}$

Proposition 3.1. *The collection* $\{\phi_i, \psi_j \mid i, j = 1, \dots, m\}$ *defined in* (3) *is a biorthogonal system, i.e.,*

$$[\phi_i, \psi_j] = \delta_{ij}, \quad i, j = 1, \dots, m.$$

Moreover, the operator $P : \mathcal{B} \to \mathbb{B}$, *defined by*

$$Pu = \sum_{i=1}^{m} [\phi_i, u]\psi_i \tag{4}$$

is the $\langle \cdot, \cdot \rangle-$ *orthogonal projection onto* $Q\mathcal{L}$. $P^* : \mathcal{B}^* \to \mathcal{B}^*$, *defined by*

$$P^*\varphi = \sum_{i=1}^{m} [\varphi, Q\phi_i] Q^{-1}\psi_i$$

is the $\langle \cdot, \cdot \rangle_*-$*orthogonal projection on* $\mathcal{L}$. *In addition,* P^* *is the adjoint of* P *in the sense that*

$$[\varphi, P\psi] = [P^*\varphi, \psi], \quad \varphi \in B^*, \psi \in \mathcal{B}$$

and we have

$$P^* = Q^{-1}PQ.$$

Proof. (I) The biorthogonality of the collection follows straightforwardly. First note that from Equation (1) the following holds

$$\begin{aligned} \langle Q\phi_i, Q\phi_j \rangle &= [Q^{-1}(Q\phi_i), Q\phi_j] \\ &= [\phi_i, Q\phi_j] \\ &= \Theta_{ij}. \end{aligned} \tag{5}$$

So, using above result Equation (5) and the symmetry of the duality product, the biorthogonality follows as

$$\begin{aligned} [\phi_i, \psi_j] &= \langle Q\phi_i, \psi_j \rangle \\ &= \langle Q\phi_i, \Big(\sum_{l=1}^{m} (\Theta^{-1})_{jl} Q\phi_l \Big) \rangle \\ &= \sum_{l=1}^{m} (\Theta^{-1})_{jl} \langle Q\phi_i, Q\phi_l \rangle \\ &= \sum_{l=1}^{m} (\Theta^{-1})_{jl} \Theta_{li} \\ &= \left(\Theta^{-1}\Theta\right)_{ji} \\ &= \delta_{ji} \\ &= \delta_{ij}. \end{aligned}$$

(II) Note that for each $i = 1, \dots, m$, $\psi_i \in Q\mathcal{L}$. Consequently, the range of the P lies inside $Q\mathcal{L}$ as well. Now let us fix l and consider $\psi = Q\phi_l$. Since

$$\begin{aligned} P\psi = PQ\phi_l &= \sum_{i=1}^{m}[\phi_i, Q\phi_l]\psi_i \\ &= \sum_{i=1}^{m}\Theta_{il}\Big(\sum_{k=1}^{m}(\Theta^{-1})_{ik}Q\phi_k\Big) \\ &= \sum_{i,k=1}^{m}\Theta_{il}(\Theta^{-1})_{ik}Q\phi_k \\ &= \sum_{k=1}^{m}\Big(\sum_{i=1}^{m}\Theta_{li}(\Theta^{-1})_{ik}\Big)Q\phi_k \\ &= \sum_{k=1}^{m}\left(\Theta\Theta^{-1}\right)_{lk}Q\phi_k \\ &= Q\phi_l \\ &= \psi, \end{aligned}$$

we obtain that $P\psi = \psi$. Since $Q\phi_l \in Q\mathcal{L}$ for each $l = 1, \dots, m$, it follows that $P\psi = \psi$ for $\psi \in Q\mathcal{L}$. Now suppose that ψ is orthogonal to $Q\mathcal{L}$:

$$0 = \langle Q\phi_l, \psi\rangle = [\phi_l, \psi] \quad \forall l = 1, \dots, m.$$

Then it follows that $P\psi = 0$ for all ψ orthogonal to $Q\mathcal{L}$, establishing the second assertion.

(III) We use previous observation that P and P^* are orthogonal projectors on $Q\mathcal{L}$ and $\mathcal{L}$, respectively. First consider $\varphi \in \mathcal{L}$ and $\psi \in Q\mathcal{L}$. Consequently,

$$\begin{aligned} [\varphi, P\psi] &= [\varphi, \psi] \\ &= [P^*\varphi, \psi]. \end{aligned}$$

Other cases involve when the functions are in the orthogonal complement of $\mathcal{L}$ and $Q\mathcal{L}$. So, for instance for $\varphi \in \mathcal{L}^\perp$ the relation is trivially satisfied since $\langle \varphi, \phi_l\rangle_* = 0$.

(IV) Fix $\phi \in \mathcal{B}^*$ and consider

$$\begin{aligned} QP^*\phi - PQ\phi &= \\ &= Q\Big(\sum_{i=1}^{m}[\phi, Q\phi_i]Q^{-1}\psi_i\Big) - \sum_{l=1}^{m}[\phi_l, Q\phi]\psi_l \\ &= \sum_{i=1}^{m}[\phi, Q\phi_i]\psi_i - \sum_{i=1}^{m}[\phi_l, Q\phi]\psi_l \\ &= 0, \end{aligned}$$

which follows from the symmetry of the duality $[\cdot,\cdot]$ product. □

Theorem 3.2. *We have* $Q\mathcal{L} = \operatorname{span}\{\psi_i \mid i = 1, \dots, m\}$*. Furthermore, the mapping* $v : \mathcal{B} \to \mathcal{B}$

$$v(u) = \sum_{i=1}^{m}[\phi_i, u]\psi_i$$

is the orthogonal projection of $\mathcal{B}$ *onto* $Q\mathcal{L}$ *and therefore has the variational formulation*

$$\|u - v(u)\|^2 = \inf_{\psi \in Q\mathcal{L}} \|u - \psi\|^2, \quad u \in \mathcal{B},$$

in other words,

$$v(u) = \operatorname*{arg\,min}_{\psi \in Q\mathcal{L}} \|u - \psi\|^2.$$

Proof. It follows from Proposition 3.1 that the ψ_i are the components of the projection operator (4), the orthogonal projection onto $Q\mathcal{L}$. Since $Q\mathcal{L}$ is a closed linear subspace of $\mathcal{B}$, we can apply the classical projection theorem[1], establishing that $v(u) = Pu$ and the assertion follows. □

Theorem 3.3. *Let $\psi_i \in \mathcal{B}$, $i = 1, \ldots, m$ be defined as in Equation* (3)*. The mapping $\Psi^* : \mathbb{R}^m \to \mathcal{B}$, defined by*

$$\Psi^*(y) = \sum_{i=1}^{m} y_i \psi_i, \quad y \in \mathbb{R}^m, \tag{6}$$

is an optimal minmax solution to

$$\inf_{\Psi} \sup_{u \in \mathcal{B}} \frac{\|u - \Psi(\Phi(u))\|^2}{\|u\|^2}.$$

Proof. Consider for a general solution such that

$$v(\Psi') = \sup_{u \in \mathcal{B}} \frac{\|u - \Psi'(\Phi(u))\|^2}{\|u\|^2} < \infty.$$

Then choose $u^* \in \ker \Phi$, and $u_\lambda = \lambda u$ with $\lambda > 0$:

$$\begin{aligned} \infty > v(\Psi') &\geqslant \sup_{\lambda > 0} \frac{\|\lambda u^* - \Psi'(\Phi(u^*))\|^2}{\|\lambda u^*\|^2} \\ &= \sup_{\lambda > 0} \frac{\|\lambda u^* - \Psi'(0)\|^2}{\|\lambda u^*\|^2}. \end{aligned}$$

Recall that Ψ' attains the infimum and $v(\Psi')$ is finite, so it implies $\Psi'(0) = 0$. Consequently,

$$v(\Psi') \geqslant \sup_{\lambda > 0} \frac{\|\lambda u^*\|^2}{\|\lambda u^*\|^2} = 1$$

so we conclude

$$v(\Psi') \geqslant 1, \Psi' : \mathbb{R}^m \to \mathcal{B}. \tag{7}$$

On the other hand, consider the solution given by Equation (6). By Proposition 3.1 the expression of Equation (6) is the orthogonal projection onto $Q\mathcal{L}$. Writing its action as $P_{Q\mathcal{L}} u = \Psi(\Phi(u))$, observe that

$$\begin{aligned} \sup_{u \in \mathcal{B}} \frac{\|u - \Psi(\Phi(u))\|^2}{\|u\|^2} &= \sup_{u \in \mathcal{B}} \frac{\|u - P_{Q\mathcal{L}} u\|^2}{\|u\|^2} \\ &\leqslant 1. \end{aligned}$$

So, the optimalitity of Ψ^* follows from Equation (7). □

3.3 Variational Properties

The projection coordinates ψ_i defined in Equation (3) can also be characterized via their variational properties.

Theorem 3.4. *1. For $y \in \mathbb{R}^m$, $\sum_{i=1}^m y_i \psi_i$ is the minimizer of*

$$\begin{aligned} &\min \|\psi\| \\ &\text{subject to } \psi \in \mathcal{B} \text{ and } [\phi_i, \psi] = y_j, j = 1, \ldots, m. \end{aligned} \tag{8}$$

2. For $i = 1, \ldots, m$, ψ_i is the minimizer of

$$\begin{aligned} &\min \|\psi\| \\ &\text{subject to } \psi \in \mathcal{B} \text{ and } [\phi_j, \psi] = \delta_{ij}, j = 1, \ldots, m. \end{aligned}$$

[1] The relationship between orthogonal projection and norm minimization is the classical projection theorem, see, [Lue97]: let $V \subset \mathcal{B}$ be a closed linear subspace and let $P_V : \mathcal{B} \to \mathcal{B}$ denote the orthogonal projection onto V.

Proof. 1. Take $\psi^\dagger = \sum_{i=1}^m y_i \psi_i \in \mathcal{B}$. Note that the biorthogonality in Proposition 3.1 implies that $\psi^\dagger$ is feasible. In fact,

$$\begin{aligned} [\phi_i, \psi^\dagger] &= \langle Q\phi_i, \psi^\dagger \rangle \\ &= \sum_{j=1}^m y_j \langle Q\phi_i, \psi_j \rangle \\ &= \sum_{j=1}^m y_j [\phi_i, \psi_j] \\ &= y_i. \end{aligned}$$

Pick another feasible function $\psi \in \mathcal{B}$. $\psi^\dagger = P\psi$ so from Proposition 3.1 $\psi^\dagger$ is the orthogonal projector of ψ onto $Q\mathcal{L}$. Consequently, $\psi - \psi^\dagger$ is orthogonal to $\psi^\dagger$ and

$$\|\psi\|^2 = \|\psi^\dagger\|^2 + \|\psi - \psi^\dagger\|^2.$$

So, $\psi^\dagger$ is the minimizer.

2. □

4 Kernel Flows algorithm:

As seen in the previous section, the optimal recovery of the problem 1 has the explicit form:

$$v^{\dagger} = \sum_{i,j=1}^{N} y_i A_{ij} Q \phi_j \tag{9}$$

Where $A = \Theta^{-1}$.

This form relies on the prior specification of a norm of the hilbert space $\mathcal{B}$ or equivalenty of a kernel K. In the following, we present the characteristic of a good kernel.

4.1 What is a good kernel

The method of Kernel flow is based in the premise that a kernel is good if there is no significant loss in accuracy in the prediction error if the number of data points is halved. This led to the introduction of the following loss function:

$$\rho = \frac{\|v^{\dagger} - v^{\star}\|^2}{\|v^{\dagger}\|^2} \tag{10}$$

Where $\|.\|$ is the norm associated to the RKHS $\mathcal{B}$. $v^{\star}$ is the optimal recovery open seeing half of the points.

In other words, a kernel is good when $v^{\star}$ is close $v^{\dagger}$ (measured using the RKHS norm). In this case, ρ has a small value (close to zero).

By denoting $\{s(1), ..., s(m)\}$ a selection of the m distinct elements of $\{1, ..., N\}$ (m=round(N/2)), $v^{\star}$ is the optimal recovery of $u^{\dagger}$ such as $(u^{\dagger}(x_{s(1)}) = y_{s(1)}), ..., (u^{\dagger}(x_{s(m)}) = y_{s(m)})$. Then, the explicit form of $v^{\star}$ is:

$$v^{\star} = \sum_{i,j=1}^{m} \overline{y}_i \overline{A}_{ij} Q \overline{\phi}_j \tag{11}$$

Where $\overline{y}_i = y_{s(i)}$, $\overline{A}_{ij} = A_{s(i)s(j)}$ and $\overline{\phi}_j = \phi_{s(j)}$

One can denote π the $m \times N$ sub-sampling matrix defined by $\pi_{ij} = \delta_{s(i)j}$, and then observe that:

$$v^{\star} = \sum_{i,j=1}^{N} y_i \tilde{A}_{ij} Q \phi_j$$

With $\tilde{A} = \pi^T \overline{A} \pi$ and $\overline{A} = (\pi \Theta \pi^T)^{-1}$.

Now, we aim to find an alternative formula for the loss function 10.

Theorem 4.1. *It holds true that:*
$\rho = 1 - \frac{y^T \tilde{A} y}{y^T A y}$ *and* $\rho \in [0, 1]$.

Before proving this theorem, let us prove the following proposition:

Proposition 4.2. *For* $v^{\dagger}$ *and* $v^{\star}$ *defined as in 9 and 11, we have:*
$\|v^{\dagger} - v^{\star}\|^2 = y^T A y - y^T \tilde{A} y$

Proof. (I) One can show that $\|v^\dagger\|^2 = y^T Ay$ and $\|v^\star\|^2 = \overline{y}^T\overline{A}\overline{y} = y^T\tilde{A}y$. In fact:

$$\begin{aligned}
\|v^\dagger\|^2 &= [Q^{-1}v^\dagger, v^\dagger] \\
&= [Q^{-1}\sum_{i,j=1}^{N} y_i A_{ij} Q\phi_j, \sum_{i,j=1}^{N} y_i A_{ij} Q\phi_j] \\
&= [\sum_{i,j=1}^{N} y_i A_{ij}\phi_j, \sum_{k,l=1}^{N} y_k A_{kl} Q\phi_l] \\
&= \sum_{i,j=1}^{N}\sum_{k,l=1}^{N} y_i A_{ij} y_k A_{kl} [\phi_j, Q\phi_l] \\
&= \sum_{i,j=1}^{N}\sum_{k,l=1}^{N} y_i A_{ij} y_k A_{kl} \Theta_{jl} \\
&= \sum_{i,j=1}^{N}\sum_{k=1}^{N} y_i A_{ij} y_k \sum_{l=1}^{N} A_{kl}\Theta_{lj} \\
&= \sum_{i,j=1}^{N}\sum_{k=1}^{N} y_i A_{ij} y_k (A\Theta)_{kj} \\
&= \sum_{i}^{N}\sum_{k=1}^{N} y_i (AA\Theta)_{ik} y_k \\
&= y^T Ay
\end{aligned}$$

In the same way, we can show the second equality.

(II) The orthogonal decomposition $\|v^\dagger\|^2 = \|v^\star\|^2 + \|v^\dagger - v^\star\|^2$ is obtained using the orthogonal projection on a convex closed theorem. Let recall this theorem:

Theorem 4.3. *Let consider $\mathcal{H}$ a hilbert space and C a convex closed subset of $\mathcal{H}$. It holds true that:*

$\forall x \in \mathcal{H}, \exists! m \in C$ *such as:* $d(x, C) = \inf_{z \in C} \|x - z\| = \|x - m\|$.

Moreover, we have $< x - m, y - m > \leqslant 0, \forall y \in C$.

Now, we apply this theorem to show that $< v^\star, v^\dagger - v^\star >= 0$:

Let us denote $C = \{\Psi \in \mathcal{B}, st : \Psi(x_{s(i)}) = y_{s(i)}, i \in \{1, .., m\}\}$.

C is a convex closed subset of $\mathcal{B}$. Take $x = 0 \in \mathcal{B}$.

It follows that: $\exists! m \in C$ such as: $\inf_{z \in C} \|z\| = \|m\|$.

Since $v^\star$ is the minimizer of 8 subject to the constraints $\Psi(x_{s(i)}) = y_{s(i)}$. Thus $m = v^\star$ and we obtain that:

$$< -v^\star, y - v^\star > \leqslant 0, \forall y \in C$$

Moreover, $v^\dagger \in C$ implies $< -v^\star, v^\dagger - v^\star > \leqslant 0$

And $-v^\dagger + 2v^\star \in C$ implies $< -v^\star, -v^\dagger + v^\star > \leqslant 0$

So, we conclude that: $< v^\star, v^\dagger - v^\star >= 0$

□

Then, theorem 4.1 follows simply from the proposition

4.2 The Fréchet derivative of ρ:

Let us fix y and π, then ρ can be seen as a function of A or more equivalently of Θ since $A = \Theta^{-1}$.

As seen in the previous section, the smaller the rho, the better is a kernel. Consequently, Kernel Flow method search for the optimal parameters of the kernel that minimize the loss function. In the following proposition, we compute the Fréchet derivative of ρ with respect to small perturbations of A or of Θ^{-1}.

Proposition 4.4. *1. Write $z = A^{-1}\tilde{A}y$ with $\tilde{A} = \pi^T(\pi A^{-1}\pi^T)^{-1}\pi$ defined as above. It holds true that:*

$$\rho(A + \epsilon S) = \rho(A) + \epsilon\frac{(1 - \rho(A))y^T Sy - z^T Sz}{y^T Ay} + \mathcal{O}(\epsilon^2)$$

2. And writing $\hat{y} = \Theta^{-1}y$ and $\hat{z} = \pi^T(\pi\Theta\pi^T)^{-1}\pi y$:

$$\rho(\Theta + \epsilon T) = \rho(\Theta) - \epsilon\frac{(1 - \rho(\Theta))\hat{y}^T T\hat{y} - \hat{z}^T T\hat{z}}{\hat{y}^T\Theta^{-1}\hat{y}} + \mathcal{O}(\epsilon^2)$$

Proof. Let show the first point: Using the formula of ρ 4.1, observe that:

$$\rho(A + \epsilon S) = 1 - \frac{y^T\pi^T[\pi(A + \epsilon S)^{-1}\pi^T]^{-1}\pi y}{y^T(A + \epsilon S)y}$$

Recall the approximation of the derivate of inverse matrix: $(A + \epsilon S)^{-1} = A^{-1} - \epsilon A^{-1}SA^{-1} + \mathcal{O}(\epsilon^2)$. Then:

$$\rho(A + \epsilon S) = 1 - \frac{y^T\pi^T[\pi A^{-1}\pi^T - \epsilon\pi A^{-1}SA^{-1}\pi^T]^{-1}\pi y}{y^T(A + \epsilon S)y} + \mathcal{O}(\epsilon^2)$$

Using the same approximation for:

$$[\pi A^{-1}\pi^T - \epsilon\pi A^{-1}SA^{-1}\pi^T]^{-1} = [\pi A^{-1}\pi^T]^{-1} + \epsilon[\pi A^{-1}\pi^T]^{-1}\pi A^{-1}SA^{-1}\pi^T[\pi A^{-1}\pi^T]^{-1}$$

. It follows that:

$$\begin{aligned}\rho(A + \epsilon S) &= 1 - \frac{y^T\pi^T[\pi A^{-1}\pi^T]^{-1}\pi y + \epsilon y^T\pi^T[\pi A^{-1}\pi^T]^{-1}\pi A^{-1}SA^{-1}\pi^T y^T\pi^T[\pi A^{-1}\pi^T]^{-1}\pi y}{y^T(A + \epsilon S)y}\\ &= 1 - \frac{y^T\tilde{A}y + \epsilon z^T Sz}{y^T(A + \epsilon S)y}\\ &= 1 - \frac{y^T\tilde{A}y + \epsilon z^T Sz}{y^T Ay + \epsilon y^T Sy}\end{aligned}$$

Recall the approximation $\frac{1}{y^T Ay + \epsilon y^T Sy} = \frac{1}{y^T Ay} - \epsilon\frac{y^T Sy}{(y^T Ay)^2} + \mathcal{O}(\epsilon^2)$. Then:

$$\begin{aligned}\rho(A + \epsilon S) &= 1 - (y^T\tilde{A}y + \epsilon z^T Sz)(\frac{1}{y^T Ay} - \epsilon\frac{y^T Sy}{(y^T Ay)^2} + \mathcal{O}(\epsilon^2))\\ &= \rho(A) + \epsilon\frac{(1 - \rho(A))y^T Sy - z^T Sz}{y^T Ay} + \mathcal{O}(\epsilon^2)\end{aligned}$$

The proof of the second point is identical. □

4.3 The gradient of ρ with respect to hyper-parameters of the kernel:

Let $K(x, x', W)$ be a family of kernels parametrized by $W = (W_1, \dots, W_l) \in \mathcal{W}$, where l is the number of parameters.

In kernel flow algorithm, we seek to find the parameters of the kernel that minimize the loss function ρ. In a practical manner, starting with initial values of the parameters W,we evolve W using stochastic gradient descent:

$$W \longleftarrow W - \epsilon\nabla_W\rho(W)$$

Where $\nabla_W \rho(W) = (\partial_{W_1}\rho(W), \dots, \partial_{W_l}\rho(W))^T$

Let us compute analytically the expression of $\partial_{W_i}\rho(W)$ for $i \in \{1, \dots, l\}$:

Corollary:

Write $\Theta = \Theta(W)$, $\hat{y} = \Theta^{-1}y$ and $\hat{z} = \pi^T(\pi\Theta\pi^T)^{-1}\pi y$. It holds true that:

$$\partial_{W_i}\rho(W) = -\frac{(1-\rho(W))\hat{y}^T(\partial_{W_i}\Theta(W))\hat{y} - \hat{z}^T(\partial_{W_i}\Theta(W))\hat{z}}{\hat{y}^T\Theta^{-1}\hat{y}} \tag{12}$$

Proof. Recall the approximation: $\Theta(W + \epsilon W') = \Theta(W) + \epsilon(W')^T\nabla_W\Theta(W) + \mathcal{O}(\epsilon^2)$. Then:

$$\begin{aligned}\rho(\Theta(W+\epsilon W')) &= \rho(\Theta(W) + \epsilon(W')^T\nabla_W\Theta(W)) + \mathcal{O}(\epsilon^2)\\ &= \rho(\Theta + \epsilon T) + \mathcal{O}(\epsilon^2)\end{aligned}$$

Where $T = (W')^T\nabla_W\Theta(W)$

Proposition 4.4 implies that:

$$\rho(W+\epsilon W') = \rho(W) - \epsilon\frac{(1-\rho(W))\hat{y}^T T\hat{y} - \hat{z}^T T\hat{z}}{\hat{y}^T\Theta^{-1}\hat{y}} + \mathcal{O}(\epsilon^2)$$

This equation is valid for all $W' \in \mathcal{W}$, let set $i \in \{1, \dots, l\}$,we choose $W' = (0, \dots, 0, 1, 0, \dots, 0)^T$, where the coefficient 1 is in index i. It follows that: $T = \partial_{W_i}\Theta(W)$.

Which proves the result.

□

5 Revisiting the interpolation problem from a signal model perspective

The goal of this project is learning a function from a finite number of sampled data points. Learning such a function is an ill-posed problem in the sense that a small error in sampled data may result in a large error in the resulting function. Because sampled data inevitably contain noise, the ill-posedness of these problems is unavoidable. Minimum norm interpolation and the regularization method are effective approaches to treat the ill-posedness.

5.1 Regression problem

For a single output system, the continuous time signal model can be written as

$$y = f(x) + \epsilon \tag{13}$$

For a d-dimensional system, we measure n snapshots of the state variable at time points $\{t_1, t_2, \cdots, t_n\}$, then we have the following measurement matrix:

$$X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_d \\ | & | & & | \end{pmatrix} = \begin{pmatrix} x^\top(t_1) \\ x^\top(t_2) \\ \vdots \\ x^\top(t_n) \end{pmatrix} = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_d(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_d(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_n) & x_2(t_n) & \cdots & x_d(t_n) \end{pmatrix}$$

Frequency method: linear regression and kernel regression

Bayesian method: Bayesian linear regression and Gaussian process regression

5.2 Linear regression

For the signal $Y = (y(t_1), y(t_2), \cdots, y(t_n))^\top$ to be linear in the unknown parameters, we assume

$$f(x) = \theta^\top x$$

Substituting the time points in to the equation and arrange it in to a matrix notation,

$$Y = X\theta + \epsilon$$

where $\theta \in \mathcal{R}^{d\times 1}$ is the unknown parameters, ϵ is the model error.

By using empirical risk minimization criterion, we can obtain the least square estimation:

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y \tag{14}$$

5.3 Geometrical interpretations

If we denote the columns of X by x_i, we have

$$f(x) = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} = \sum_{i=1}^{d} \theta_i x_i$$

so that the signal model is seen to be a linear combination of the "signal" vectors $\begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}$.

Then, the LS error can be written as

$$\mathcal{L}(\theta) = \|Y - \sum_{i=1}^{d} \theta_i x_i\|^2$$

We now see that the linear regression attempts to minimize the square of the distance from the data vector Y to the a signal vector $\sum_{i=1}^{d} \theta_i x_i$, which is a linear combination of the columns of X. The data vector can lie anywhere in an $N-$dimensional space, termed $\mathcal{R}^N$, while all possible signal vectors, being linear combinations of $d < N$ vectors, must lie in a d-dimensional subspace of $\mathcal{R}^N$, termed $\mathcal{X}^d$. For example, $N = 3$ and $p = 2$ we illustrate this in Figure 1. Note that all possible choices θ_1, θ_2 produce signal vectors constrained to lie in the subspace $\mathcal{X}^2$ and that in general Y does not lie in the subspace. It should be intuitively clear that the vector $\hat{x}$ that lies in $\mathcal{X}^2$ and that is closest to Y in the in the Euclidean sense is the component of Y in $\mathcal{X}^2$. Alternatively, $\hat{x}$ is the *orthogonal projection* of Y onto $\mathcal{X}^2$. This means that the error $\epsilon = Y - \hat{x}$ is orthogonal to the signal space.

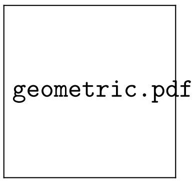

Figure 1: Geometrical viewpoint of linear regression in $\mathcal{R}^3$

In summary, the linear regression can be interpreted as the problem of fitting or approximating a data vector Y in $\mathcal{R}^N$ by another vector $\hat{x}$, which is linear combination of vectors $(x_1, x_2, \cdots, x_d)$. The problem is solved by choosing $\hat{x}$ in the subspace to be the orthogonal projection of Y.

5.4 Kernel linear regression

$$y = \phi(x)^\top \theta + \epsilon \tag{15}$$

where ϕ is a feature map: $\mathcal{R}^d \to \mathcal{R}^m$ with $m > d$.

The objective function can be written as

$$\mathcal{L}(\theta) = \|\Phi\theta - Y\|^2$$

where $\Phi = (\phi(x_1), \phi(x_2), \cdots, \phi(x_d))^\top \in \mathcal{R}^{n \times d}$.

Calculating the derivative the objective function and set to be zero gives the optimal solution

$$\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top Y \tag{16}$$

Lemma 5.1. $(\Phi^\top \Phi)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top)^{-1}$. *This is can be shown by using the singular value decomposition.*

let $x^\star$ be a test data,in the light of Lemma 5.1, the optimal prediction can be obtained

$$f(x^\star) = \phi(x^\star)^\top \hat{\theta} = \phi(x)^\top \Phi^\top (\Phi\Phi^\top)^{-1} Y \tag{17}$$

Note that $[\Phi\Phi^\top]_{ij} = \phi(x(t_i))^\top \phi(x(t_j)) := K(x(t_i), x(t_j))$, and $[\phi(x^\star)^\top \Phi^\top]_j = \phi(x^\star)^\top \phi(x(t_j)) := K(x^\star, x(t_j))$. Therefore, the equation (17) can be written as

$$f(x^\star) = K(x^\star, X) k(X, X)^{-1} Y \tag{18}$$

Representer theorem: In statistical learning theory, a representer theorem is any of several related results stating that a minimizer $f^\star$ of a regularized empirical risk functional defined over a reproducing kernel Hilbert space can be represented as a finite linear combination of kernel products evaluated on the input points in the training set data.

Theorem 5.2. *Consider a positive-definite real-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on a non-empty set $\mathcal{X}$ with a corresponding reproducing kernel Hilbert space $\mathcal{H}_k$. The approximation f in the kernelized form can be express as*

$$f(x) = \sum_{j=1}^{m} \alpha_j K(x(t_j), x) \tag{19}$$

Representer theorem can dramatically reduce an infinite dimensional problem to a dimensional one whose solution can be obtained by solving either a linear system or a finite dimensional optimization problem.

5.5 Bayesian linear regression

From Bayesian perspective, we assume that θ is a random variable whose particular realization we must estimate. Consider linear model

$$\begin{aligned} f(x) &= \theta^\top x \\ y &= f(x) + \epsilon \\ \epsilon &\sim (0, \sigma^2) \end{aligned}$$

Build the Bayesian model:

Introduce Gaussian prior $p(\theta) = \mathcal{N}(0, \Sigma_p)$, then the posterior $p(\theta|Data)$ can be obtained by

$$p(\theta|X, Y) = \frac{p(\theta, Y|X)}{p(Y|X)} = \frac{p(Y|\theta, X)p(\theta|X)}{\int p(Y|\theta, X)p(\theta|X)d\theta}$$

Note that denominator is dependent with the parameter and $p(\theta|X) = p(\theta)$, therefore,

$$p(\theta|X, Y) \propto p(Y|\theta, X)p(\theta) = \prod_{i=1}^{N} p(Y(t_i)|x(t_i), \theta) \cdot \mathcal{N}(0, \Sigma_p)$$

According to the signal model, we have $p(Y(t_i)|x(t_i), \theta) = \mathcal{N}(\theta^\top x(t_i), \sigma^2)$, we have

$$p(\theta|X, Y) \propto \prod_{i=1}^{N} \mathcal{N}(\theta^\top x(t_i), \sigma^2) \cdot \mathcal{N}(0, \Sigma_p)$$

Therefore, the problem becomes

$$p(\theta|Data) \sim \mathcal{N}(\mu_\theta, \Sigma_\theta) \propto \prod_{i=1}^{N} \mathcal{N}(\theta^\top x(t_i), \sigma^2) \cdot \mathcal{N}(0, \Sigma_p) \tag{20}$$

Inference:

First, calculate the likelihood

$$\begin{aligned} \prod_{i=1}^{N} \mathcal{N}(\theta^\top x(t_i), \sigma^2) &= \frac{1}{(2\pi)^{N/2}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y(t_i) - \theta^\top x(t_i))^2\right) \\ &= \frac{1}{(2\pi)^{N/2}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(Y^\top - \theta^\top X^\top)(Y - X\theta)\right) \\ &= \frac{1}{(2\pi)^{N/2}\sigma^2} \exp\left(-\frac{1}{2}(Y^\top - \theta^\top X^\top)\sigma^{-2}I(Y - X\theta)\right) \end{aligned}$$

Substituting the result into equation (20) and after a simple manipulation, we have

$$\begin{aligned} p(\theta|Data) &\propto \exp\left(-\frac{1}{2}(Y^\top - \theta^\top X^\top)\sigma^{-2}I(Y - X\theta) - \frac{1}{2}\theta^\top \Sigma_p^{-1}\theta\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}(Y^\top Y - 2Y^\top X\theta + \theta^\top X^\top X\theta) - \frac{1}{2}\theta^\top \Sigma_p^{-1}\theta\right) \end{aligned}$$

Recall a exponential part of a pdf $p(x) \sim \mathcal{N}(\mu, \Sigma)$ is

$$\exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) = \exp\left(-\frac{1}{2}(x^\top \Sigma^{-1} x - 2\mu^\top \Sigma^{-1} x + c)\right)$$

Comparing the two function gives the quadratic and first terms, respectively

$$x^\top \Sigma^{-1} x = \theta^\top \sigma^{-2} X^\top X \theta + \theta^\top \Sigma_p^{-1} \theta = \theta^\top (\sigma^2 X^\top X + \Sigma_p^{-1}) \theta \tag{21}$$

where $A := \Sigma_\theta^{-1} = \sigma^2 X^\top X + \Sigma_p^{-1}$.

$$\mu^\top \Sigma^{-1} x = \sigma^{-2} Y^\top X \theta \tag{22}$$

which means $\mu_\theta^\top \Sigma_\theta^{-1} = \sigma^{-2} Y^\top X \Rightarrow \mu_\theta = \sigma^{-2} A^{-1} X^\top Y$.

Prediction:

Given a $x^\star$, the predicted result is

$$f(x^\star) = x^{\star\top} \theta \sim \mathcal{N}(x^{\star\top} \mu_\theta, x^{\star\top} \Sigma_\theta x^\star)$$

Recall $Y = f(x) + \epsilon$, so we have the final prediction

$$p(Y^\star | Data, x^\star) \sim \mathcal{N}(x^{\star\top} \mu_\theta, x^{\star\top} \Sigma_\theta x^\star + \sigma^2 I) \tag{23}$$

5.6 Gaussian process regression

Weight space view

Recall that the linear scenario $p(f(x)|Data, x^\star) \sim \mathcal{N}(x^{\star\top} \sigma^{-2} A^{-1} X^\top Y, x^{\star\top} A^{-1} x^\star)$. If we define a nonlinear feature mapping like equation (15), so the pdf is

$$p(f(x^\star)|Data, x^\star) \sim \mathcal{N}(\phi(x^\star)^\top \sigma^{-2} A^{-1} \Phi(X)^\top Y, \phi(x^\star)^\top A^{-1} \phi(x^\star)) \tag{24}$$

By using *Woodbury Matrix Identity*

$$(A_{n\times n} + U_{n\times k} C_{k\times k} V_{k\times n})^{-1} = A^{-1} - A^{-1} U (C^{-1} + V A^{-1} U)^{-1} V A^{-1}$$

equation (24) can be manipulated as

$$p(f(x^\star)|Data, x^\star) \sim \mathcal{N}(\phi(x^\star)^\top \Sigma_p \Phi (\Phi \Sigma_p \Phi^\top + \sigma^2 I)^{-1} Y, \phi(x^\star)^\top \Sigma_p \phi(x^\star) - \phi(x^\star)^\top \Sigma_p \Phi^\top (\Phi \Sigma_p \Phi^\top + \sigma^2 I)^{-1} \Phi \Sigma_p \phi(x^\star))$$

Because Σ_p is a symmetric positive definite matrix, let $\Sigma_p = (\Sigma_p^{\frac{1}{2}})^2$, then we can get

$$\begin{aligned} K(x, x') &= \phi(x)^\top \Sigma_p^{\frac{1}{2}} \Sigma_p^{\frac{1}{2}} \phi(x') \\ &= (\Sigma_p^{\frac{1}{2}} \phi(x))^\top \cdot \phi(x) \Sigma_p^{\frac{1}{2}} \\ &= < \varphi(x, \varphi(x')) > \end{aligned} \tag{25}$$

which is also called 'kernel trick'. Finally, the posterior is

$$p(f(x^\star)|Data, x^\star) \sim \mathcal{N}(K(x^\star, X)(K(X, X) + \sigma^2 I)^{-1} Y, K(X, X) - K(x^\star, X)(K(X, X) + \sigma^2 I)^{-1} K(X, x^\star)) \tag{26}$$

How to link to Gaussian process?

For a sequence of random variables $\{X_t\}_{t\in T}$, where T is a continuous domain. If and only if for every finite set of indices $t_1, \cdots, t_k$ in the index set T, $X_{t_1, \cdots, t_k}$ is a multivariate Gaussian random variable, then $\{X_t\}_{t\in T}$ is a Gaussian process.

Given a prior $p(\theta) \sim \mathcal{N}(0, \Sigma_p)$, the expectation of $f(x)$ is

$$E[f(x)] = E[\phi(x)^\top \theta] = \phi(x)^\top E[\theta] = 0$$

For $\forall x, x' \in \mathcal{R}^d$,

$$\begin{aligned} cov(f(x), f(x')) &= E[(f(x) - E[f(x)])(f(x') - E[f(x')])] \\ &= E[f(x)f(x')] \\ &= E[\phi(x)^\top \theta \phi(x')^\top \theta] \\ &= E[\phi(x)^\top \theta \theta^\top \phi(x')] \\ &= \phi(x)^\top E[\theta\theta^\top]\phi(x') \\ &= \phi(x)^\top E[(\theta - 0)(\theta^\top - 0)]\phi(x') \\ &= \phi(x)^\top \Sigma_p \phi(x') \\ &= K(x, x') \end{aligned}$$

Obviously, the covariance of $f(x)$ is a kernel function.

Function space view

In this regression problem, $f(x)$ is a Gaussian process.

$$\begin{aligned} &\{f(x)\}_{x \in \mathcal{R}^p} \sim GP(m(x), K(x, x')) \\ &m(x) = E[f(x)], \quad K(x, x') = E[(f(x) - m(x))(f(x') - m(x'))^\top] \end{aligned} \tag{27}$$

The regression problem is $y = f(x) + \epsilon \sim \mathcal{N}(\mu(X), K(X, X) + \sigma^2 I)$

Given a new data: $X^\star = (x_1^\star, x_2^\star, \cdots, x_N^\star)_{N \times d}$, the adjoint probability density is

$$\begin{bmatrix} Y \\ f(x^\star) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu(x) \\ \mu(x^\star) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X^\star) \\ K(X^\star, X) & K(X^\star, X^\star) \end{bmatrix}\right)$$

Next we calculate $p(f(X^\star)|Y, X, X^\star)$, i.e $p(f(X^\star)|Y)$, a conditional PDF. Recall that the equation

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix} = \left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$$

$$x_b | x_a \sim \mathcal{N}(\mu_{b|a}, \Sigma_{b|a})$$

$$\mu_{b|a} = \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a) + \mu_b$$

$$\Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$$

using this equation gives the result

$$\begin{aligned} p(f(X^\star)|Y) = &\mathcal{N}(K(x^\star, X)(K(X, X) + \sigma^2 I)^{-1}(Y - \mu(X)) + \mu(X^\star), \\ &K(X, X) - K(X^\star, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X^\star)) \end{aligned} \tag{28}$$

Add noise

$$\begin{aligned} p(f(X^\star)|Y) = &\mathcal{N}(K(x^\star, X)(K(X, X) + \sigma^2 I)^{-1}(Y - \mu(X)) + \mu(X^\star), \\ &K(X, X) - K(X^\star, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X^\star) + \sigma^2 I) \end{aligned}$$

In conclusion, in the GPR method,

1. From the weight space perspective, the prediction is

$$p(y^\star | Data, x^\star) = \int p(y^\star | x^\star, \theta) p(\theta) d\theta$$

2. From the function space perspective, the prediction is

$$p(y^\star | Data, x^\star) = \int p(y^\star | x^\star, f(x)) p(f(x)) df(x)$$

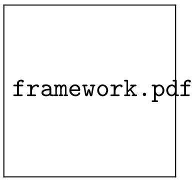

Figure 2: Relationship between the four methods

References

[Lue97] David G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, NY, USA, 1997.

[OS19] Houman Owhadi and Clint Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2019.

[OY19] Houman Owhadi and Gene Ryan Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.